

A Neural Text Ranking Approach for Automatic MeSH Indexing

Alastair R. Rae, James G. Mork and Dina Demner-Fushman

National Library of Medicine, 8600 Rockville Pike, Bethesda, MD, 20894, USA

Abstract

The U.S. National Library of Medicine (NLM) has been indexing the biomedical literature with MeSH terms since the mid-1960s, and in recent years the library has increasingly relied on AI assistance and automation to curate the biomedical literature more efficiently. Since 2002, the NLM has been using natural language processing algorithms to assist indexers by providing MeSH term recommendations, and we are continually working to improve the quality of these recommendations. This work presents a new neural text ranking approach for automatic MeSH indexing. The domain-specific pretrained transformer model, PubMedBERT, was fine-tuned on MEDLINE data and used to rank candidate main headings obtained from a Convolutional Neural Network (CNN). Pointwise, listwise, and multi-stage ranking approaches are demonstrated, and the algorithm performance was evaluated by participating in the BioASQ challenge task 9a on semantic indexing. The neural text ranking approach was found to have very competitive performance in the final batch of the challenge, and the multi-stage ranking method typically boosted the CNN model performance by about 5% points in terms of micro F1-score.

Keywords

Automatic MeSH Indexing, Medical Text Indexing, Neural Text Ranking, Transformers, BERT

1. Introduction

The U.S. National Library of Medicine (NLM) maintains the MEDLINE[®] bibliographic database to help the biomedical research community find the journal articles that they need. To improve the quality of PubMed search results, all MEDLINE articles are indexed with a controlled vocabulary called Medical Subject Headings (MeSH[®])¹.

MeSH indexing is a time-consuming and highly specialized activity. NLM indexers review the full text of an article and then assign MeSH terms that represent the central concepts as well as every other topic that is discussed to a significant extent. This work focuses on the indexing of main headings, which are also known as MeSH descriptors. There are currently over 29,000 main headings in the 2021 MeSH vocabulary and each main heading describes an important biomedical concept. On average indexers assign about 11 main headings per article.


Each year close to 1 million articles are indexed for MEDLINE, and the library uses AI assistance and automation to increase the efficiency of the indexing process. Since 2002, indexing assistance has been provided by the Medical Text Indexer (MTI) system[1]. MTI

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ alastair.rae@nih.gov (A. R. Rae); jmork@mail.nih.gov (J. G. Mork); ddemner@mail.nih.gov (D. Demner-Fushman)



© 2021 No copyright. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.nlm.nih.gov/mesh/>

improves productivity by providing a pick list of recommended MeSH terms that can be quickly selected by indexers.

Automatic MeSH indexing is a difficult machine learning problem, and the main challenges are the large number of main headings and their highly imbalanced frequency distribution. This work presents a new neural text ranking approach for automatic MeSH indexing. Pointwise, listwise, and multi-stage ranking approaches are demonstrated using a domain-specific pre-trained transformer model called PubMedBERT[2]. To the best of our knowledge this is the first time that text ranking using pretrained transformers has been applied to the automatic MeSH indexing problem. The performance of the new approach was evaluated by participating in the BioASQ challenge task 9a on semantic indexing[3].

2. Related Work

In recent years two high-performing approaches for automatic MeSH indexing have emerged: learning-to-rank approaches, and neural network multi-label classification approaches. Examples of learning-to-rank based systems are MTI[1] and DeepMeSH[4], and examples of neural network multi-label classification systems are MeSHProbeNet[5] and AttentionMeSH[6]. Recently, You et al. achieved state-of-the-art performance with BERTMeSH[7]. BERTMeSH is a neural network multi-label classification approach that leverages pretrained transformers[8] and the article full-text.

Learning-to-rank[9] is a methodology that uses supervised machine learning algorithms to solve ranking problems. Typically, it is applied to automatic MeSH indexing by treating the title and abstract as the query, and candidate main headings as the documents to be ranked. Learning-to-rank algorithms usually make use of hand-crafted features and rank documents by integrating multiple sources of evidence. For example, MTI uses a learning-to-rank algorithm[10] to rank candidate main headings from MetaMap[11], PubMed Related Citations[12], and machine learning algorithms. Sources of evidence include text features such as the fraction of main heading unigrams and bigrams that appear in the title or abstract.

Learning-to-rank algorithms can be classified as pointwise, pairwise, and listwise approaches depending on the loss function that is used. Pointwise approaches compute a loss for individual query-document pairs. The training task is to predict whether individual candidate documents are relevant to a query. At inference time, the model is run on all query-document pairs, and overall rankings are obtained using the predicted relevance scores. Pairwise approaches compute a loss for a query and a pair of documents. The training task is to predict which document is more relevant to a query. At inference time, for each query, pairwise rankings are obtained for all candidate documents pairs, and then these pairwise rankings are converted into an overall ranking. Finally, listwise approaches compute a loss for a query and all candidate documents. The training task is to predict the correct overall document ranking for a query. Hence, listwise approaches directly solve the ranking problem.

Recently, neural text ranking using pretrained transformers has proven to be a very effective approach for ad-hoc information retrieval[13]. On the MS MARCO passage ranking dataset large-scale pretrained transformer models, such as BERT[14], have outperformed traditional information retrieval approaches by a considerable margin[15]. Text ranking using pretrained

transformers was first demonstrated by Nogueira and Cho[16]. They implemented a pointwise approach by training BERT as a relevance classifier on MS MARCO query-passage pairs. Pairwise text ranking using BERT was also demonstrated as part of a multi-stage ranking architecture[17]. To the best of our knowledge there is no prior work on listwise text ranking using pretrained transformers. For a recent review of text ranking with pretrained transformers the interested reader is referred to “Pretrained Transformers for Text Ranking: BERT and Beyond” by J. Lin et al.[13].

Domain-specific pretraining of transformer models can improve performance on downstream tasks[8, 2], and BioBERT[8] is a popular domain-specific version of BERT that has been pre-trained on a biomedical corpus. The BioBERT authors started with the original BERT checkpoint and ran additional pretraining steps on a corpus of PubMed abstracts and PubMed Central article full-text. Recently, PubMedBERT was shown to outperform BioBERT on the Biomedical Language Understanding and Reasoning Benchmark (BLURB)[2]. PubMedBERT was also pre-trained on PubMed abstracts and PubMed Central article full-text, however, unlike BioBERT, it was trained from scratch using a domain-specific vocabulary.

3. Methods

This section describes our automatic MeSH indexing approaches that were evaluated by participating in the BioASQ challenge task 9a on semantic indexing.

3.1. Convolutional Neural Network

The baseline approach was our previously described Convolutional Neural Network (CNN) for automatic MeSH indexing[18]. It is a neural network multi-label classification approach that takes the article title, abstract, journal, publication year, and indexing year as input. The top N results from this model were also used as candidate main headings for the text ranking approaches.

3.2. Pointwise Text Ranking

The neural text ranking approaches were implemented using a domain-specific pretrained transformer model called PubMedBERT[2]. PubMedBERT is a BERT model with a domain-specific vocabulary that has been pretrained from scratch on a biomedical corpus. It was chosen because it was the top performing model in the BLURB benchmark[2], and also because its domain-specific vocabulary was expected to encode biomedical text efficiently. More details about the BERT architecture and fine-tuning configurations can be found in the original BERT paper[14].

For the pointwise text ranking approach PubMedBERT was configured as a relevance classifier using the text pair classification configuration. The input sequence was:

$$[[CLS], q, [SEP], d, [SEP]], \quad (1)$$

where the query, q , comprises the concatenated tokens of the indexing year, journal name, title, and abstract, and d comprises the tokens of the candidate main heading. $[CLS]$ and $[SEP]$ are

the classification and separator special tokens respectively.

In the text pair classification configuration, the $[CLS]$ token is used to represent the input sequence, and the relevance probability was computed by adding a softmax classification head on top of its contextualized embedding ($T_{[CLS]}$):

$$P(\text{Relevant} = 1 | d_i, q) = s_i \triangleq \text{softmax}(T_{[CLS]}W + b)_1, \quad (2)$$

where W and b are the weights and bias of the classification layer respectively, and $\text{softmax}(\cdot)_i$ denotes the i -th element of the softmax output.

The training task was to predict whether a candidate main heading is relevant, given the article title, abstract, and other metadata. For the automatic MeSH indexing task, a main heading is considered relevant if it was indexed. PubMedBERT was fine-tuned on positive and negative article-candidate main heading pairs sampled from the CNN top results using the cross-entropy loss:

$$L = - \sum_{j \in J_{pos}} \log(s_j) - \sum_{j \in J_{neg}} \log(1 - s_j), \quad (3)$$

where J_{pos} is a set of indexes for article-main heading pairs where the main heading was indexed, and J_{neg} is a set indexes for article-main heading pairs where the main heading was not indexed.

At inference time, the fine-tuned model was run on all candidate main headings from the CNN top results and predicted relevance scores were used to generate a per-article main heading ranking. The final set of predicted main headings for an article was obtained by applying a decision threshold to the ranking scores.

3.3. Listwise Text Ranking

For the listwise text ranking approach PubMedBERT was configured for text tagging, and the second text input was the shuffled candidate main headings separated by the pipe symbol. The input sequence was therefore:

$$[[CLS], q, [SEP], |, d_1, |, d_2, \dots, |, d_N, [SEP]]. \quad (4)$$

The pipe symbols allow the model to distinguish between different candidate main headings, and random shuffling was employed to prevent overfitting to the main heading order.

Relevance probabilities were computed by feeding the contextualized embedding of the first token of each main heading to the softmax classification head described in Equation 2. Thus, the listwise approach directly creates a ranking by performing relevance classification on all candidate main headings at once. PubMedBERT was fine-tuned on the top N main headings from the CNN model using cross-entropy loss. Again, the final set of main headings was obtained by applying a decision threshold.

3.4. Multi-Stage Text Ranking

The listwise approach is expected to outperform the pointwise approach because it can consider interactions between main headings. However, a problem with the listwise approach is that the

length of the input sequence is proportional to the number of candidate main headings, and this limits the recall of the approach because there is a maximum number of candidate main headings that can fit within BERT's maximum sequence length of 512 tokens. The multi-stage text ranking approach attempts to overcome this limitation by first ranking the candidate main headings using the pointwise approach. The pointwise approach can rank any number of main headings, and it is expected to have higher recall@N than the CNN model. For the multi-stage ranking approach it was also found to be beneficial to average the ranking scores of the different stages.

More formally, starting with the CNN ranking, R_0 , the top N_p results were reranked using the pointwise approach to generate a ranking R_1 . R_0 and R_1 scores were then averaged to generate R_2 . Next, the top N_l results in R_2 were reranked using the listwise approach to generate ranking R_3 . The final ranking was computed by averaging the scores of R_2 and R_3 . A decision threshold was applied to generate the final main heading predictions.

3.5. Multi-Stage Text Ranking with COVID-19 Rules

During the BioASQ challenge it was noticed that our machine learning approaches were performing poorly on COVID-19 articles, and two specific problems were identified:

- The "COVID-19" main heading was always being indexed with unnecessary additional main headings "Pneumonia, Viral", "Coronavirus Infections", and "Pandemics".
- The "SARS-CoV-2" main heading was always being indexed with the unnecessary additional main heading "Betacoronavirus".

The precision of the multi-stage text ranking approach was improved by removing these unnecessary main headings using manually written COVID-19 rules.

3.6. Hybrid Approach

MTI First Line Indexing (MTIFL) and MTI Review Filtering (MTIR) are used for selected journals to partially automate the NLM indexing process². For MTIFL journals, MTI provides the initial indexing, and this is later reviewed (and potentially modified) by human indexers. For MTIR journals the process is the same except that human curation is only used for critical elements.

Empirically, MTI is found to perform very well for semi-automatically indexed journals. In order to achieve the highest possible agreement with human indexing, we therefore implemented a hybrid approach that used MTI First Line Index results for MTIFL journals, Default MTI results for MTIR journals, and multi-stage text ranking results (with COVID-19 rules) for all other journals. Default MTI is configured for balanced precision and recall, and its results were expected to be the most similar to the initial indexing provided for MTIR journals. MTI results are made publicly available for anyone wanting to use them during the challenge.

²<https://ii.nlm.nih.gov/MTI/MTIFL.shtml>

4. Experiments

4.1. BioASQ Task 9a

The performance of our proposed approaches were evaluated by participating in the large-scale online biomedical semantic indexing task of the 2021 BioASQ challenge (task 9a). The task released 3 batches of 5 test sets, and these contained between 4,000 and 11,000 soon-to-be-indexed MEDLINE articles. Participants had a limited time window to submit results, and this was necessary to ensure that indexing predictions were made before indexer annotations became available.

The NLM team used the challenge to evaluate various different text ranking approaches and configurations. This paper describes our final best-performing approaches, and these are evaluated on the 5 weekly test sets of batch 3. Results for our final approaches were only submitted to the last two test sets of batch 3, and for a more comprehensive performance evaluation, we have independently generated indexing predictions for the week 1-3 test sets.

4.1.1. Dataset

The dataset was constructed from the MEDLINE/PubMed 2021 annual baseline³. Fully and semi-automatically indexed articles (“Automated” or “Curated” indexing method) were excluded as we believe that the indexing of these articles may be biased by MTT’s predictions. 20,000 randomly selected articles published in 2020/2021 were reserved for a validation set, and another 40,000 randomly selected articles published in 2020/2021 were reserved for our personal test set. The remaining 10 million articles published after 2006 were used for the training set.

The presented approaches were evaluated on the BioASQ task 9a batch 3 test sets, and independently generated predictions were evaluated using indexer annotations downloaded from the NLM E-Utilities⁴ service on the 28th of June 2021. The final challenge results were calculated using the indexing available on the 21st of May 2021, and to allow for fair comparisons between systems, indexing completed after this date was excluded.

4.1.2. Evaluation Metrics

The primary evaluation metric used by the semantic indexing task is the micro F1-score (MiF) and this is defined as the harmonic mean of the micro precision (MiP) and the micro recall (MiR):

$$MiF = \frac{2 \cdot MiP \cdot MiR}{MiP + MiR}, \quad (5)$$

where

$$MiP = \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N_L} y_{ij} \cdot \hat{y}_{ij}}{\sum_{i=1}^{N_A} \sum_{j=1}^{N_L} \hat{y}_{ij}}, \quad (6)$$

$$MiR = \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N_L} y_{ij} \cdot \hat{y}_{ij}}{\sum_{i=1}^{N_A} \sum_{j=1}^{N_L} y_{ij}}. \quad (7)$$

³https://www.nlm.nih.gov/databases/download/pubmed_medline.html

⁴<https://www.ncbi.nlm.nih.gov/books/NBK25497/>

In the above equations y are the indexer annotations, \hat{y} are the model predictions, N_A is the number of articles, and N_L is the number of main headings. Model predictions were made after applying a decision threshold to the predicted scores. There is an optimum decision threshold that results in the highest F1-score, and this threshold was determined by a linear search on the validation set.

4.1.3. Configuration

The configuration for the CNN model has previously been described in Rae et al.[18], and the model was retrained on the MEDLINE/PubMed 2021 dataset described in this paper.

The pointwise and listwise ranking models were implemented using the Hugging Face Transformers library (v4.2.2) with a PyTorch (v1.7.1) backend. PubMedBERT pretrained weights were downloaded from the Hugging Face model repository, and the uncased model pretrained on abstracts and full-text was selected (“BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext”).

The pointwise model was implemented in Hugging Face Transformers using the BertForSequenceClassification class (specifying the number of labels as 2), and the default PubMedBERT configuration was left unchanged. Training was run for approximately 1 epoch on a balanced dataset, and the Adam optimizer was used with L2 weight decay set to 0.01. The learning rate schedule included 10,000 warmup steps, a maximum learning rate of 2e-5, and a linear decay to zero thereafter.

The listwise model was implemented in Hugging Face Transformers using the BertForTokenClassification class, with the number of labels set to 2. All tokens, except for the first token of each main heading, were assigned the masking label of -100. The first token of each main heading was assigned a label of 1 or 0 for indexed and not-indexed main headings respectively. Again, the PubMedBERT configuration was not altered, and the model was trained on the CNN top 50 results for approximately 10 epochs. Other training settings were the same as for the pointwise approach, except that a lower maximum learning rate of 9e-6 was used.

Both ranking models were trained on the Biowulf cluster⁵ using NVIDIA V100x 32GB GPUs. The pointwise and listwise models were trained on 4 and 2 GPUs respectively for approximately 10 days. FP16 training was used and an effective batch size of 128 was achieved using gradient accumulation. Validation set performance of the listwise model had converged after 10 days, however the performance of the pointwise model was still improving.

For the hybrid approach, MTI results⁶ and MTIFL and MTIR journal lists⁷ (22nd of September 2020 versions) were downloaded from the NLM website.

4.1.4. Results

Table 1 summarizes the micro F1-score performance of top performing systems in batch 3. For each weekly test set, the table includes the highest micro F1-score achieved by each team, along with the best performing MTI baseline for reference. The table shows that the performance of

⁵<https://hpc.nih.gov/>

⁶<http://ii.nlm.nih.gov/BioASQ/>

⁷<https://ii.nlm.nih.gov/MTI/MTIFL.shtml>

Table 1

Micro F1-score performance of top performing systems in batch 3.

System	Week 1	Week 2	Week 3	Week 4	Week 5	Average
NLM System 3 (<i>hybrid approach</i>)	0.7059	0.6973	0.6966	0.6999	0.7075	0.7014
dmiip_fdu systems	0.7060	0.6976	0.6980	0.6966	0.7013	0.6999
MTI First Line Index	0.6555	0.6445	0.6541	0.6491	0.6508	0.6508
pi_dna	0.6443	0.6464	0.6503	0.6466	0.6498	0.6475
DeepSys2	0.5780	0.5674		0.5651	0.5625	0.5683
iria-1	0.4895	0.4778	0.4758	0.4818	0.4729	0.4796

Table 2

Micro F1-score performance of NLM approaches in batch 3.

Approach	Week 1	Week 2	Week 3	Week 4	Week 5	Average
Hybrid (<i>NLM System 3</i>)	0.7059	0.6973	0.6966	0.6999	0.7075	0.7014
Multi-stage + COVID-19 rules	0.7032	0.6971	0.6953	0.6932	0.7011	0.6980
Multi-stage (<i>NLM System 2</i>)	0.7000	0.6945	0.6931	0.6894	0.6932	0.6940
Listwise (<i>NLM System 4</i>)	0.6931	0.6888	0.6884	0.6836	0.6876	0.6883
Pointwise (<i>NLM System 1</i>)	0.6888	0.6831	0.6820	0.6801	0.6799	0.6828
CNN (<i>NLM CNN</i>)	0.6482	0.6434	0.6424	0.6381	0.6424	0.6429

the neural text ranking approach is very competitive. Our best performing hybrid approach outperformed the MTI baseline by about 5% points, and it has very similar performance to the state-of-the-art dmiip_fdu systems.

Table 2 shows micro F1-score performance of NLM approaches in batch 3. Note that results for the “Multi-stage + COVID-19 rules” approach were not submitted to the challenge because teams were allowed a maximum of 5 systems. Comparing the performance of the multi-stage ranking approach to the CNN model, it can be seen that neural text ranking provided about a 5% point performance boost on average. The table shows that the listwise approach outperformed the pointwise approach and also that multi-stage ranking was beneficial. The COVID-19 rules provided small but consistent performance improvements, and the hybrid approach, which substituted MTI results for semi-automatically indexed journals, was the best performing NLM system in all batch 3 test sets.

4.2. Listwise Model Hyperparameter Search

There is a performance trade-off for the listwise approach: increasing the number of candidate main headings (N) increases the maximum achievable recall, but it also results in more input truncation due to longer input sequence lengths. This section explores this trade-off through a hyperparameter search for the optimum number of candidate main headings for the listwise approach.

For the study, the listwise model was trained with four different values of N between 25 and 50, and input truncation percentages and model performance were measured on the validation set. BioBERT input truncation percentages were also measured for comparison.

The results of the study are shown in Figure 1. Figure 1a shows a significant increase in

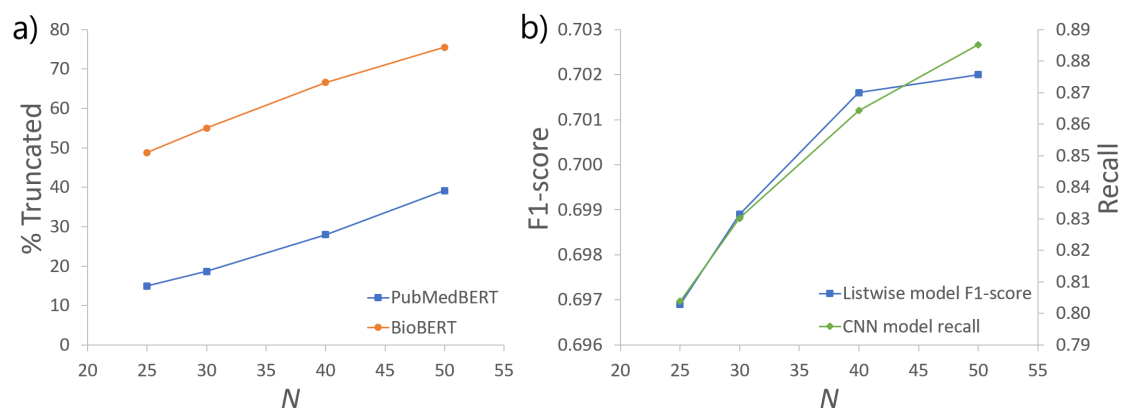


Figure 1: a) Percentage of truncated inputs vs. number of candidate main headings (N) for PubMedBERT and BioBERT. b) Listwise model micro F1-score and CNN model recall vs. number of candidate main headings.

the percentage of truncated inputs as the number of candidate main headings is increased from 25 to 50. For PubMedBERT, 15% of inputs were truncated for 25 candidate main headings, and this rises to 39% of inputs for 50 candidate main headings. The figure also shows that input truncation percentages were much higher for BioBERT than for PubMedBERT, and this is because BioBERT does not have a domain-specific vocabulary.

Despite the relatively high input truncation percentages observed in Figure 1a, Figure 1b shows that the listwise model micro F1-score increases with N , and the model trained with 50 candidate main headings is shown to have the highest micro F1-score of 0.7020 on the validation set. As expected, the increase in micro F1-score is correlated with the increase in CNN model recall, but for $N = 50$ the strength of this correlation appears to be weakening.

5. Discussion

As expected, the results indicate that the listwise text ranking approach outperforms the pointwise text ranking approach, but to confirm this we would need to train the pointwise model to convergence and also optimize the number of candidate main headings. The pointwise approach considers one article-main heading pair per training example, whereas the listwise approach considers 50 article-main heading pairs per training example, and so it makes sense that training of the pointwise model would converge more slowly.

This work has presented a hyperparameter search for the optimum number of candidate main headings for the listwise approach and increasing the number of candidate main headings from 25 to 50 was shown to result in a 0.51% point improvement in micro F1-score performance on the validation set. Increasing the number of candidate main headings further may result in additional performance improvements, however; for $N = 50$ there is some evidence that input truncation is starting to limit performance. The study also indicates that PubMedBERT was a good model choice because it was shown to encode biomedical text more efficiently than

BioBERT resulting in significantly less input truncation. For 50 candidate main headings the BioBERT tokenizer was shown to truncate about 75% of input sequences, and this would likely have a large negative impact on MeSH indexing performance.

The poor performance of our machine learning models on COVID-19 articles (before applying the COVID-19 rules) can be explained by inconsistent and out-of-date training data. The problem is that indexing of COVID-19 articles has evolved during the pandemic due to changing indexing rules and also after the addition of COVID-19 specific main headings. This is an interesting example of how sudden data and concept drift have been problematic for machine learning systems during the COVID-19 pandemic.

Finally, substituting MTI predictions for semi-automatically indexed journals was shown to consistently improve performance. An explanation could be that indexing of MTIFL and MTIR journals is biased by MTI's predictions.

6. Conclusion

This paper has presented a new neural text ranking approach for automatic MeSH indexing. PubMedBERT was fine-tuned on MEDLINE data and used to rank candidate main headings obtained from a CNN model. Pointwise, listwise, and multi-stage text ranking approaches were demonstrated, and their performance was evaluated on batch 3 of the BioASQ 2021 semantic indexing task. The neural text ranking approach was shown to have very competitive performance, and the multi-stage text ranking method was found to boost the CNN model micro F1-score performance by about 5% points.

In the future, we would like to investigate the zero-shot performance of neural text ranking models for automatic MeSH indexing. In particular, it would be interesting to know if they can correctly index a new main heading for a concept that has only been seen during unsupervised pretraining. It would be very useful if the text ranking models are learning the general concept of “indexing relevance” rather than specific indexing rules for each main heading.

Acknowledgments

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

- [1] J. Mork, A. Aronson, D. Demner-Fushman, 12 years on - is the NLM medical text indexer still useful and relevant?, *J. Biomed. Semant.* 8 (2017) 8.
- [2] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2021. [arXiv:2007.15779](https://arxiv.org/abs/2007.15779).
- [3] M. Krallinger, A. Krithara, A. Nentidis, G. Paliouras, M. Villegas, BioASQ at CLEF2020: Large-scale biomedical semantic indexing and question answering, in: *Advances in Information Retrieval*, Springer International Publishing, 2020, pp. 550–556.

- [4] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, S. Zhu, DeepMeSH: deep semantic representation for improving large-scale MeSH indexing, *Bioinformatics* 32 (2016) i70–i79.
- [5] G. Xun, K. Jha, Y. Yuan, Y. Wang, A. Zhang, MeSHProbeNet: a self-attentive probe net for MeSH indexing, *Bioinformatics* (2019).
- [6] Q. Jin, B. Dhingra, W. Cohen, X. Lu, AttentionMeSH: simple, effective and interpretable automatic MeSH indexer, in: 6th BioASQ Workshop, Brussels, Belgium, 1 November 2018. Proceedings of the 6th BioASQ Workshop, ACL, 2018, pp. 47–56.
- [7] R. You, Y. Liu, H. Mamitsuka, S. Zhu, BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text, *Bioinformatics* 37 (2020) 684–692.
- [8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2019) 1234–1240.
- [9] T.-Y. Liu, Learning to rank for information retrieval, *Found. Trends Inf. Retr.* 3 (2009) 225–331.
- [10] I. Zavorin, J. Mork, D. Demner-Fushman, Using learning-to-rank to enhance NLM medical text indexer results, in: 4th BioASQ workshop, Berlin, Germany, 12–13 August 2016. Proceedings of the Fourth BioASQ workshop, ACL, 2016, pp. 8–15.
- [11] A. R. Aronson, F.-M. Lang, An overview of metamap: historical perspective and recent advances, *J. Am. Med. Inform. Assoc.* 17 (2010) 229–236.
- [12] J. Lin, J. W. Wilbur, PubMed related articles: a probabilistic topic-based model for content similarity, *BMC Bioinformatics* 8 (2007) 423.
- [13] J. Lin, R. Nogueira, A. Yates, Pretrained transformers for text ranking: BERT and beyond, 2020. [arXiv:2010.06467](https://arxiv.org/abs/2010.06467).
- [14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), ACL, 2019, pp. 4171–4186.
- [15] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, 2021. [arXiv:2102.07662](https://arxiv.org/abs/2102.07662).
- [16] R. Nogueira, K. Cho, Passage re-ranking with BERT, 2020. [arXiv:1901.04085](https://arxiv.org/abs/1901.04085).
- [17] R. Nogueira, W. Yang, K. Cho, J. Lin, Multi-stage document ranking with BERT, 2019. [arXiv:1910.14424](https://arxiv.org/abs/1910.14424).
- [18] A. R. Rae, D. O. Pritchard, J. G. Mork, D. Demner-Fushman, Automatic mesh indexing: Revisiting the subheading attachment problem, in: AMIA 2020, American Medical Informatics Association Annual Symposium, Virtual Event, USA, November 14–18, 2020, AMIA, 2020.